야구 기사와 선수 기록을 활용한 딥러닝 기반 극성 판별 모델 개발

저자
(Authors)

박영준, 김형석, 김동화, 이한규, 김보섭, 류나현, 김성범, 강필성

출처
(Source)

대한산업공학회 춘계공동학술대회 논문집 , 2016.4, 3126-3141(16 pages)

발행처
(Publisher)

대한산업공학회
Korean Institute Of Industrial Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06673104

APA Style

박영준, 김형석, 김동화, 이한규, 김보섭, 류나현, 김성범, 강필성 (2016). 야구 기사와 선수 기록을 활용한 딥러닝 기반 극성 판별 모델 개발. 대한산업공학회 춘계공동학술대회 논문집, 3126-3141

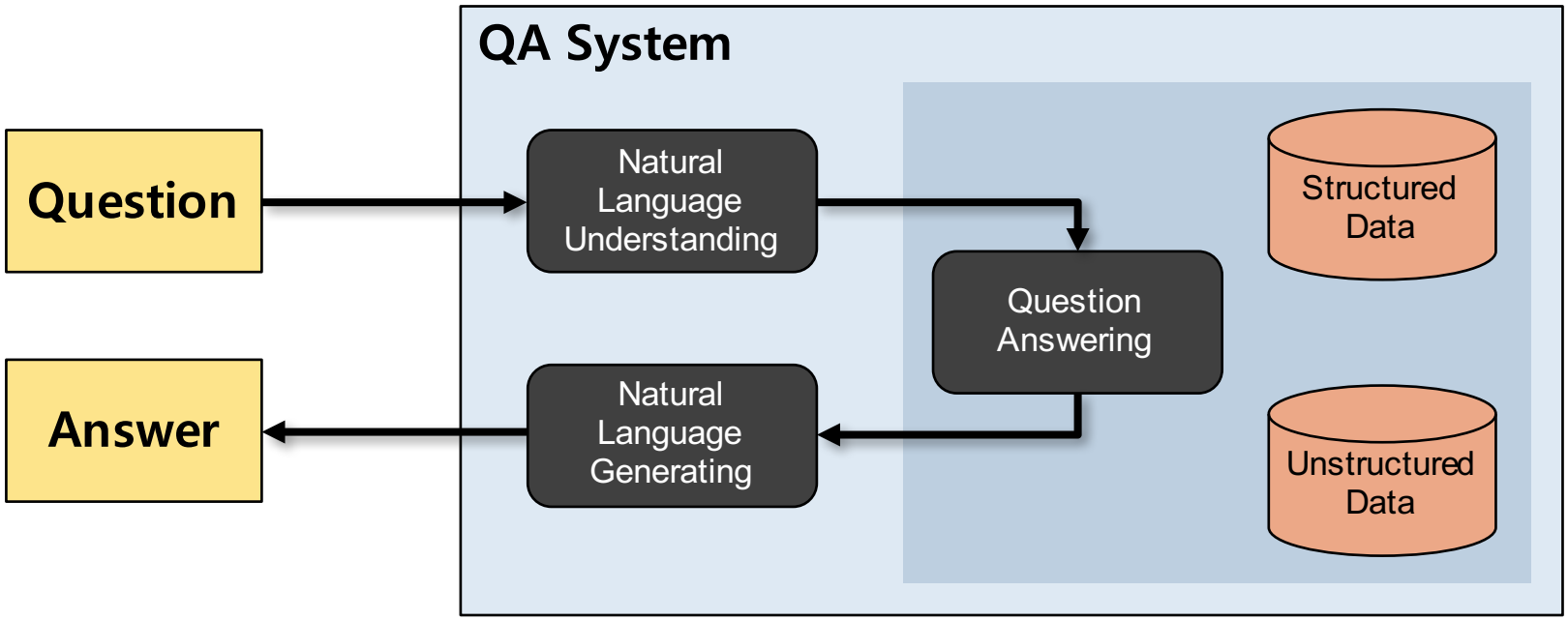
야구 기사와 선수 기록을 활용한 딥러닝 기반 극성 판별 모델 개발

박영준, 김형석, 김동화, 이한규, 김보섭, 류나현, 김성범, 강필성
고려대학교 산업공학과

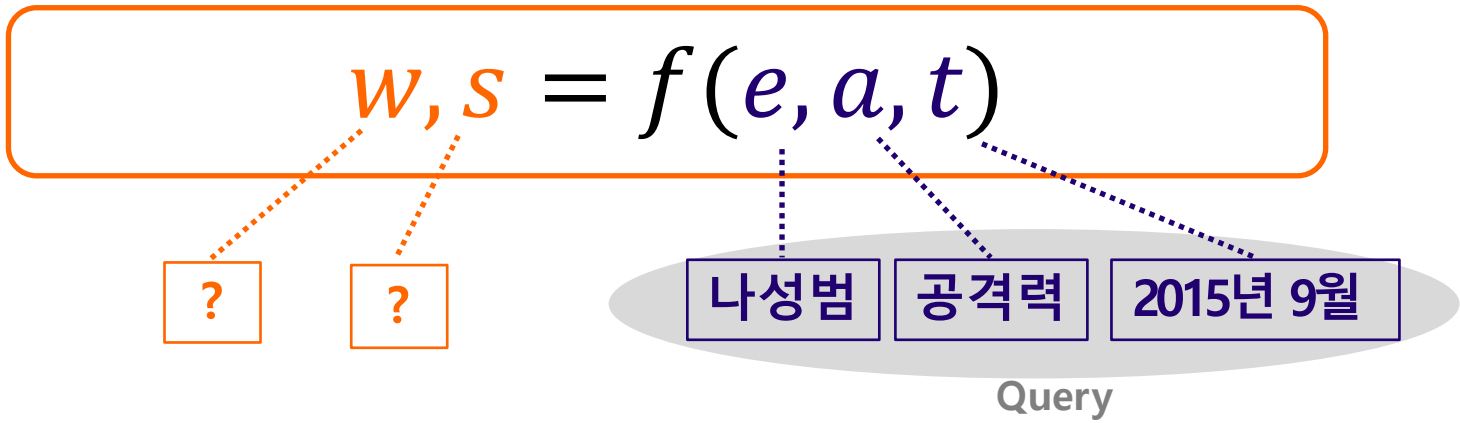
목차

연구배경: 질의응답 시스템

- 자연어 형식의 질문에 대해 질문의 분야(domain)에 적합한 답변을 제공해줄 수 있는 질의-응답 방법론에 대한 개발
- 자연어 처리, 정형/비정형의 데이터 기계학습/텍스트마이닝 등의 요소가 결합된 시스템



연구목표

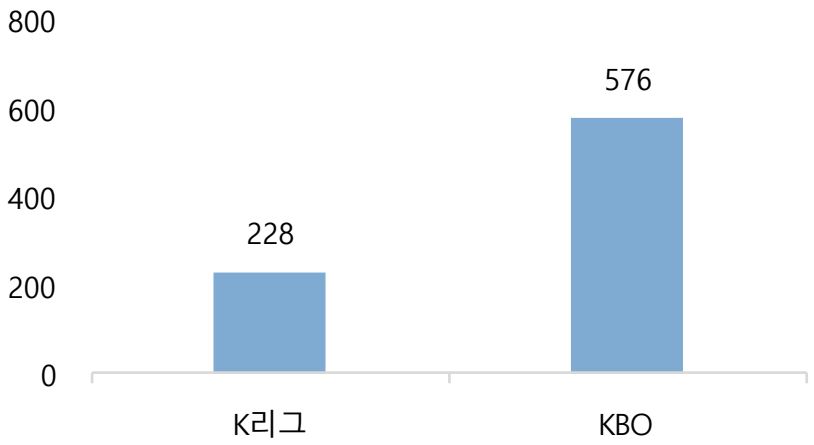


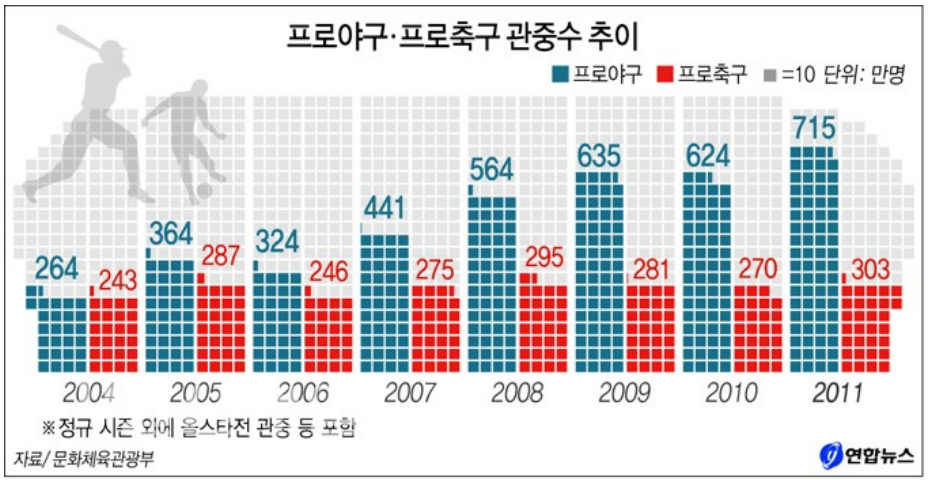
- $e \in E$: Entity in finite entity space (ex: 나성범, 한화 이글스, NC vs. 한화)
- $a \in A$ or $\varnothing$: Attribute in finite attribute space (ex: 공격력, 성적, 재미)*
- $t \in T$: Target date of inference (ex: 최근, 작년, 어제)
- $\boldsymbol{w} \in \boldsymbol{V}$: Word expressions in vocabulary (ex: 좋다, 훌륭하다, 흥미롭다)
- $\boldsymbol{s} \in \mathbb{R}$: Confidence score, represents the certainty of a tuple

한국 프로야구

- 상대적으로 높은 국민 관심도
- 9개 구단, 매년 총 576경기
- 2010 - 2014년, Nate 야구기사
- 2010 - 2014년 선수 개인기록

KBO LEAGUE





관련연구

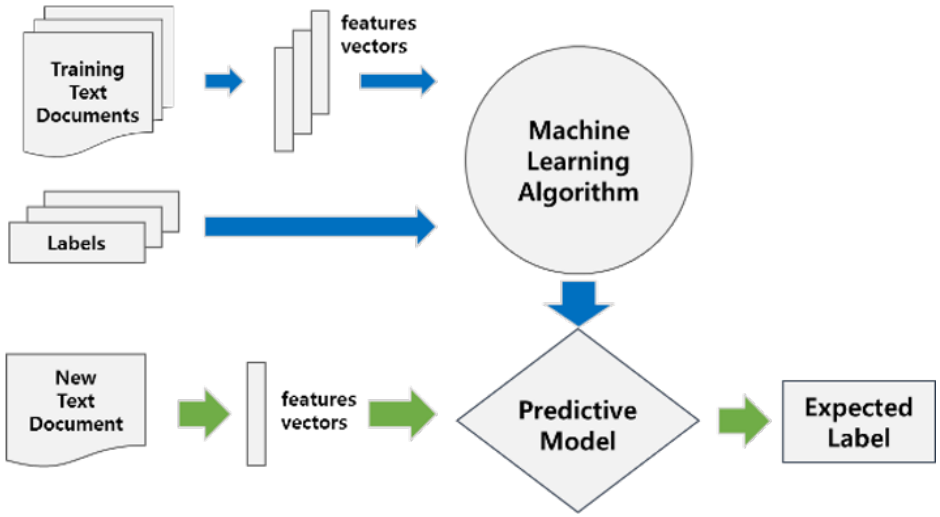
- Sentiment Classification for Text Data
- Sentiment Classification depend on Ontology

Sentiment Classification for Text Data

- Lexicon-based sentiment analysis
  - 감성어 사전 **필요**, 도메인에 **종속**
- Corpus-based sentiment analysis
  - 감성어 사전 **불필요**, 도메인에 **독립**
  - Deep learning 모델 이용하여 성능 향상

This camera is great[+1].	+1 (Positive)
I found it beautiful[+1] and good[+1].	+2 (Positive)
It looks terrible[-1].	-1 (Negative)
This car has blue color.	0 (Neutral)

<Lexicon-based Method>



<Corpus-based Method>

Sentiment Classification depend on Ontology

- 평가하고 하는 대상의 속성이 여러개 일때, 온톨로지 이용
- 온톨로지 구성에 사람이 개입
- 고비용, 편향된 결과

디자인: 외관, 화려하다, 유려한

디스플레이: 액정, 선명한, 어둡다

프로세서: 속도, 빠르다

⋮

방[illegible] 노히트, 완봉

위기관리[illegible] 브, 강판

제구력[illegible], 실

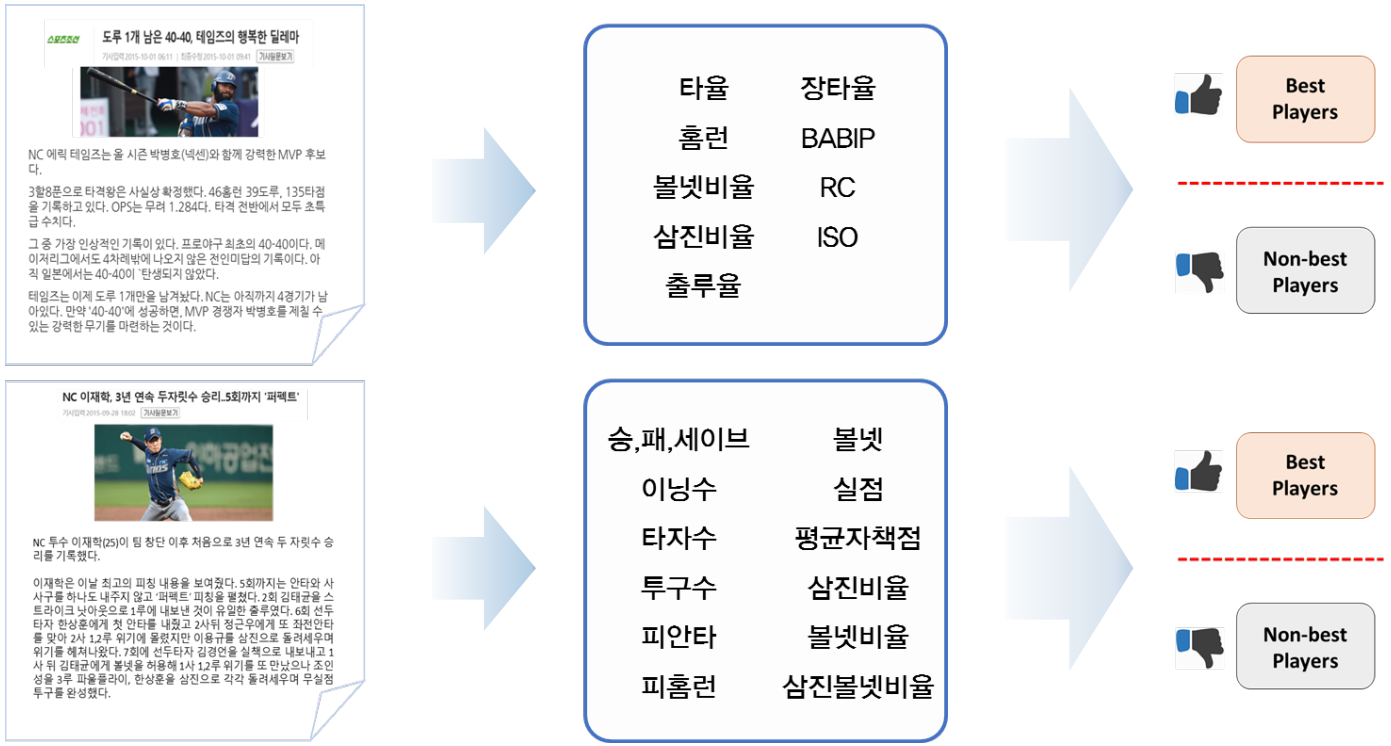
⋮

제안 방법론

- Heterogeneous Deep Learning Model
  - “나성범 잘 하니?”
  - “어떤 근거로?”
- Label Annotation for Unlabeled Sentence
  - 뉴스 기사에는 명시적인 레이블이 없음
  - ‘**잘 한다**’는 것에 대한 정량적인 기준이 없음

Heterogeneous Deep Learning Model

- Multiclass prediction + best/non-best player prediction
- Multiclass prediction: X=뉴스기사, Y=스탯변화
- Best/non-best player prediction: X=스탯변화, Y=베스트 플레이어 여부



Label Annotation for Unlabeled Sentence

- 전일 대비 선수 스탯의 변화량

Best Player Case

테임즈 첫 그랜드슬램 폭발, 시즌 11호 홈런
OSEN | 2014.05.29. | 네이버뉴스
NC 외국인 타자 에릭 테임즈가 첫 만루 홈런을 쏘아올렸다. 테임즈는 29일 대전구장에서 열린 2014 한국야쿠르트 세븐 프로야구 한화와 원정경기에 5번타자 1루수로 선발출장, 5-3으로 리드한 4회 2사 만루에서...

<2014.05.29 NC 테임즈 관련기사>

이전 경기기록

날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-28	테임즈	3	2	0	2	0.315	0	1	0

볼넷 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당 홈런	RC	BABIP
0.33	0	Inf	0.75	1.00	1.75	0.68	0	2.25	0.67

당일 경기기록

날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-29	테임즈	**6**	**5**	**7**	**4**	**0.333**	**2**	0	0

볼넷 비율	삼진 비율	볼/삼 비율	**출루율**	**장타율**	**ops**	**ISO**	**타석당 홈런**	**RC**	**BABIP**
0	0	NA	**0.83**	**2.00**	**2.83**	**1.67**	**0.333**	**10.00**	**0.75**

Target

타수	안타	타점	득점	타율	홈런	타율	홈런	출루율	장타율	ops	ISO	타석당 홈런	RC	BABIP
+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1

볼넷	삼진	볼넷	볼넷 비율	삼진 비율	볼/삼 비율
0	0	0	0	0	0

Best player 1

Non-best Player Case

'프로 12년차' 손시헌, 통산 1000경기 출장 눈앞
파이낸셜뉴스 | 2014.05.29. | 네이버뉴스
손시헌(34,NC)이 통산 1,000경기 출장을 눈앞에 뒀다. 지난 2003년 두산에서 선수생활을 시작한 손시헌은 군 입대로 인해 자리를 비운 2007년과 2008년을 제외하고 지난해까지 두산에서 뛰었다. 두산 유니폼을 입는...

<2014.05.29 NC 손시헌 관련기사>

이전 경기기록

날짜	선수	타수	안타	타점	득점	타율	홈런	볼넷	삼진
05-28	손시헌	4	1	0	0	0.292	0	0	0

볼넷 비율	삼진 비율	볼/삼 비율	출루율	장타율	ops	ISO	타석당 홈런	RC	BABIP
0	0	0	0.25	0.25	0.5	-0.042	0	0.25	0.25

당일 경기기록

날짜	선수	타수	안타	타점	**득점**	타율	홈런	**볼넷**	삼진
05-29	손시헌	4	1	0	**1**	0.291	0	**1**	0

볼넷 비율	삼진 비율	**볼/삼 비율**	**출루율**	장타율	**ops**	**ISO**	**타석당 홈런**	**RC**	**BABIP**
0.25	0	**Inf**	**0.4**	0.25	**0.65**	**-0.041**	**0**	**0.40**	**0.25**

Target

득점	볼넷	볼넷 비율	출루율	ops	ISO	타석당 홈런	RC	BABIP
+1	+1	+1	+1	+1	+1	+1	+1	+1

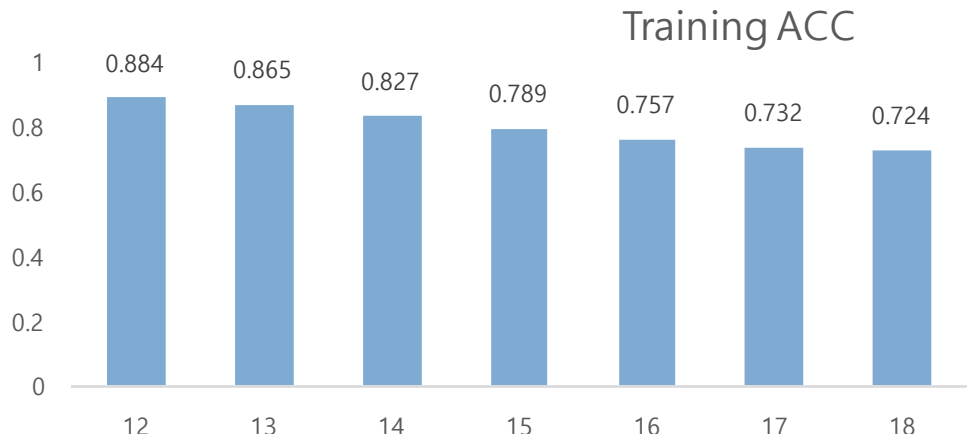
타수	안타	타점	타율	홈런	삼진	삼진 비율	장타율
0	0	0	0	0	0	0	0

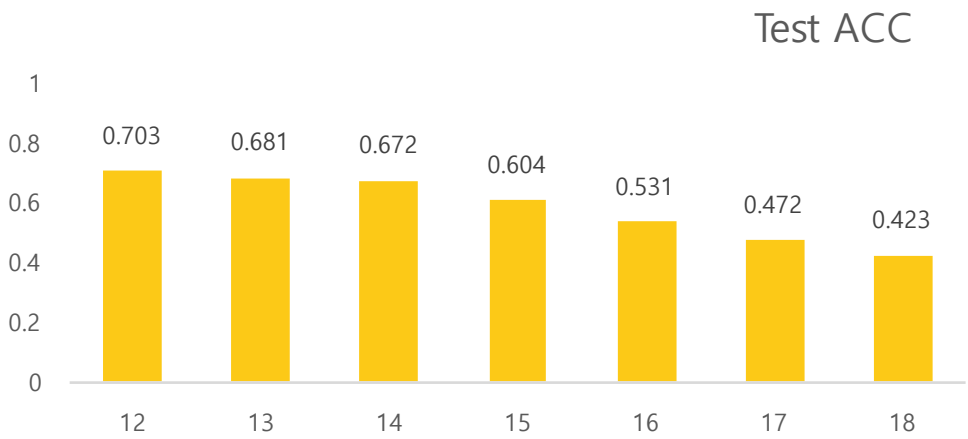
Best player 0

Experiments: 투수

- Multiclass prediction

P(x ≥ n)	Training ACC	Test ACC
…	…	…
12개 이상	0.884	0.703
13개 이상	0.865	0.681
14개 이상	0.827	0.672
15개 이상	0.789	0.604
16개 이상	0.757	0.531
17개 이상	0.732	0.472
18개 모두	0.724	0.423

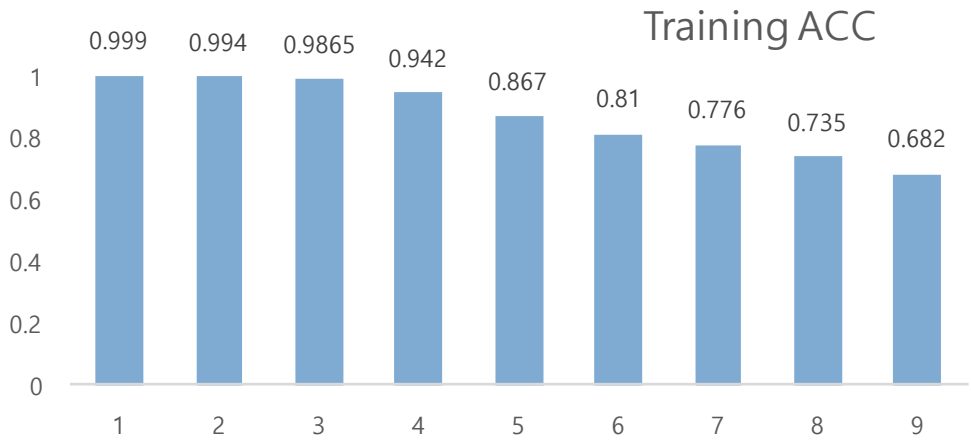


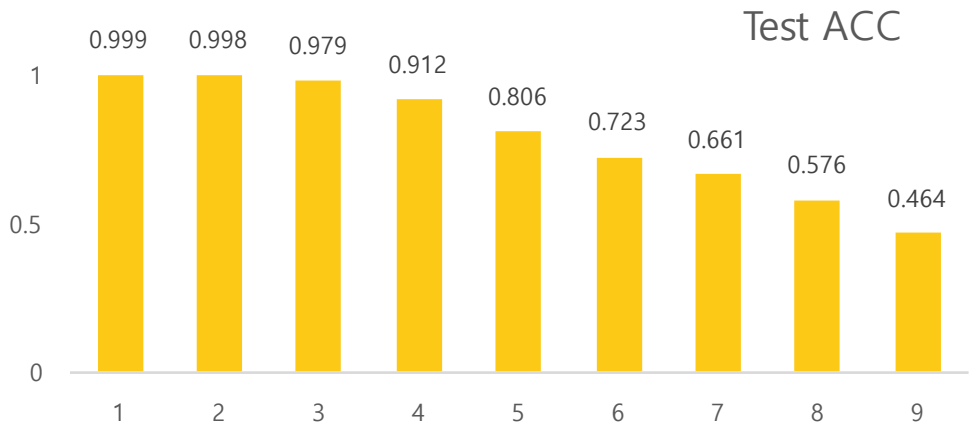


Experiments: 타자

- Multiclass prediction

P(x ≥ n)	Training ACC	Test ACC
1개 이상	0.999	0.999
2개 이상	0.994	0.998
3개 이상	0.987	0.979
4개 이상	0.942	0.912
5개 이상	0.867	0.806
6개 이상	0.810	0.723
7개 이상	0.776	0.661
8개 이상	0.735	0.576
9개 모두	0.682	0.464





실험 결과

- Best/non-best player prediction

- 투수의 경우

18개 기록을 통한 학습 모델 결과

실제 기록	예측 기록	실제 기록 수	예측된 기록 수	백분율
Best player	Best player	193	171	**88.60%**
	Non-best Player		22	11.40%
Non-best Player	Best player	1,285	87	6.77%
	Non-best Player		1,198	**93.23%**
Total		1,478	1,369	**92.62%**

- 타자의 경우

9개 기록을 통한 학습 모델 결과

실제 기록	예측 기록	실제 기록 수	예측된 기록 수	백분율
Best player	Best player	208	174	**83.65%**
	Non-best Player		34	16.35%
Non-best Player	Best player	537	137	25.51%
	Non-best Player		400	**74.49%**
Total		745	574	**77.05%**

실험결과: 최종

- Heterogeneous deep learning model

- 투수의 경우

  예측된 18개 기록을 통한
  베스트 플레이어의 분류 결과

실제 문장	예측 문장	실제 문장 수	예측된 문장 수	백분율
Best player	Best player	14,314	11,358	**79.35%**
	Non-best Player		2,956	20.65%
Non-best Player	Best player	5,686	1,624	28.56%
	Non-best Player		4,062	**71.44%**
Total		20,000	15,420	**77.1%**

- 타자의 경우

  예측된 9개 기록을 통한
  베스트 플레이어의 분류 결과

실제 문장	예측 문장	실제 문장 수	예측된 문장 수	백분율
Best player	Best player	13,817	11,168	**80.83%**
	Non-best Player		2,649	19.17%
Non-best Player	Best player	6,183	2,468	39.92%
	Non-best Player		3,715	**60.08%**
Total		20,000	14,883	**74.42%**

결론

- 정형 데이터와 비정형 데이터를 이용한 질의 응답 시스템의 구성 방법론 제안
- 딥러닝의 구조적 특성을 이용하여 이형의 데이터로 학습한 두 개의 모델을 결합
- 한국 프로야구 데이터와 관련 기사를 통하여 제안하는 방법론을 검증
- 향후 feed forward neural network 이외에 text를 모델링 하는데 더 좋은 성능을 나타내는 recurrent neural network를 적용
- 제안 방법론을 다른 분야로 확장적용 및 평가